Am. J. Hum. Genet. 47:499–514, 1990

Application of DNA Fingerprints for Cell-Line Individualization

Dennis A. Gilbert,*,† Yvonne A. Reid,‡ Mitchell H. Gail,§ David Pee,‖ Christine White,‡ Robert J. Hay,‡ and Stephen J. O'Brien*

*Laboratory of Viral Carcinogenesis, National Cancer Institute, Frederick, MD; †Department of Biology, Johns Hopkins University, Baltimore; ‡American Type Culture Collection, and §Biostatistics Branch, National Cancer Institute, Rockville, MD; and ‖Information Management Services, Silver Spring, MD

Summary

DNA fingerprints of 46 human cell lines were derived using minisatellite probes for hypervariable genetic loci. The incidence of 121 *Hae*III DNA fragments among 33 cell lines derived from unrelated individuals was used to estimate allelic and genotypic frequencies for each fragment and for composite individual DNA fingerprints. We present a quantitative estimate of the extent of genetic difference between individuals, an estimate based on the percentage of restriction fragments at which they differ. The average percent difference (APD) among pairwise combinations from the population of 33 unrelated cell lines was 76.9%, compared with the APD in band sharing among cell lines derived from the same individual (≤1.2%). Included in this survey were nine additional cell lines previously implicated as HeLa cell derivatives, and these lines were clearly confirmed as such by DNA fingerprints (APD ≤ 0.6%). On the basis of fragment frequencies in the tested cell line population, a simple genetic model was developed to estimate the frequencies of each DNA fingerprint in the population. The median incidence was 2.9×10^{-17}, and the range was 2.4×10^{-21} to 6.6×10^{-15}. This value approximates the probability that a second cell line selected at random from unrelated individuals will match a given DNA fingerprint. Related calculations address the chance that any two DNA fingerprints would be identical among a large group of cell lines. This estimate is still very slight; for example, the chance of two or more common DNA fingerprints among 1 million distinct individuals is <.001. The procedure provides a straightforward, easily interpreted, and statistically robust method for identification and individualization of human cells.

Introduction

The development of in-vitro propagation of human cells has been an extraordinary technical advance which has applications in virtually all phases of human biology. The history of this technology, however, has been somewhat tainted since its inception by a serious incidence of mistaken cell identity and subsequent invalid conclusions due to cell contamination (Gartler 1967, 1968; Nelson-Rees et al. 1974, 1981; Lavappa et al. 1976; Nelson-Rees and Flandermeyer 1976; O'Brien et al. 1980). The most common contaminant was the notorious HeLa cell line, which is an aggressive cervical carcinoma cell line established by George Gey from a tumor growing in a 31-year-old black woman, Henrietta Lacks, who died from the condition in 1951 (Gey 1956; Jones et al. 1971).

The first suggestion of frequent cell mix up came in 1967 when 18 human cell cultures were tested for an isozyme marker, *G6PD,* and all expressed a $G6PD^A$ phenotype, normally restricted to blacks (Gartler 1967, 1968). Since most of the indicated cell lines were purportedly derived from Caucasian patients, it appeared as if each had been taken over by an aggressive human tumor line, HeLa, which was also $G6PD^A$. These conclusions were strongly affirmed when common unique chromosome translocation "markers" in each of the suspect cell lines were found to be identical to those originally described in prototype HeLa culture (Nelson-Rees et al. 1974). During the next decade over 90 HeLa cell contaminations of human cells masquerading under

Received July 3, 1989; final revision received April 19, 1990.

Address for correspondence and reprints: Dr. Stephen O'Brien, Laboratory of Viral Carcinogenesis, National Cancer Institute, Frederick, MD 21701.

different names were uncovered (Nelson-Rees et al. 1974, 1980, 1981; Nelson-Rees and Flandermeyer 1976; Harris et al. 1981). At the time, these HeLa contaminants represented over one-third of the human fibroblast cell lines developed for cancer research and cell biology. The financial loss, in man-hours and reagents, from the HeLa contaminants is incalculable but is likely to be in the tens of millions of dollars (Gold 1986).

Nelson-Rees's detective work employed G-trypsin karyotyping and allozyme typing for *G6PD,* procedures diagnostic for HeLa contamination but not necessarily informative for cell contamination by other human cell lines. For such cases, *HLA* typing (Ferrone et al. 1971; Hsu et al. 1976) has been employed as well as the use of six polymorphic gene-enzyme systems termed the "allozyme genetic signature" (O'Brien et al. 1977, 1980). The reliability of *HLA* typing is limited by occasional loss of antigens as well as by unscheduled induction of novel *HLA* epitopes during cell-culture passage (Mann et al. 1983; Collins et al. 1986). The allozyme genetic signature is statistically powerful for comparing two or three cell lines, because it is based on the cells' composite allozyme genotype frequency in human populations, always $\leq .02$ (O'Brien et al. 1980). However, when many cell lines of a collection are typed for genetic uniqueness, the probability of chance identity increases. For example, in a sample of 20 cell lines the expected probability of a match by chance is nearly .50 (Gail et al. 1979; O'Brien et al. 1980).

With the advances of gene cloning, a number of genetically hypervariable regions have been discovered in human DNA (Wyman and White 1980; Bell et al. 1982; Proudfoot et al. 1982; Reeders et al. 1985; Stoker et al. 1985). These regions, which are often members of related gene families dispersed to several human chromosomes, are composed of tandem repeats of short core sequences that display abundant variation in the number of repeat units among alleles and among loci (Jeffreys et al. 1985*b;* Nakamura et al. 1987).

Jeffreys and co-workers have described a set of DNA probes from such tandem repetitive minisatellites which on hybridization to genomic DNA detect multiple hypervariable loci producing a DNA "fingerprint," a complex Southern blot pattern of multiple polymorphic DNA fragments derived from numerous hypervariable loci (Jeffreys et al. 1985*b*). DNA fingerprints show germline stability and, with the exception of MZ twins, are completely individual specific (Jeffreys et al. 1985*c*). They have been used successfully in forensic medicine (Gill et al. 1985), in paternity determination for both human and animal populations (Jeffreys et al. 1985*a;* Burke and Bruford 1987; Jeffreys and Morton 1987; Wetton et al. 1987; Lynch 1988), and in preliminary reports for cell-line analyses (Thacker et al. 1988; van Helden et al. 1988; Mann et al. 1989).

We report here the application of DNA fingerprinting to a population of 46 human cell lines. The high level of polymorphism detected by these probes allowed an unambiguous individualization of each cell line tested. Included in this analysis were nine cell lines which had been previously implicated as HeLa derivatives by cytogenetic and allozyme screens (for review, see O'Brien et al. 1977, 1980; Nelson-Rees et al. 1980, 1981). These cell lines showed a DNA fingerprint nearly band-for-band identical to the original HeLa reference stock. A parameter which quantitates the amount of genetic difference between individuals, termed "average percent difference" (APD), was calculated for each of 1,035 cell-line pair combinations, based upon the occurrence of 121 unique bands in this population. DNA fragment frequency was related to allele frequency at multiple ($N \cong 10$) homologous genetic loci by using a simple genetic model that allowed the estimation of the probability of any particular DNA fingerprint, a value exceedingly small, in most cases $< 1 \times 10^{-16}$. The application of these techniques to a population of cultured human cells provides both a straightforward and robust measure of a cell-line individuality (or not) and certain population genetic parameters of the alleles tracked by the minisatellite probes.

Material and Methods

Cell Lines

All human cell lines were obtained from the American Type Culture Collection, Rockville, MD. The HeLa cells and cells with HeLa marker chromosomes and related characteristics were CCL 2 (HeLa), CCL 5 (L-132), CCL 6 (Intestine 407), CCL 13 (Chang Liver), CCL 17 (KB), CCL 18 (Detroit 98), CCL 21 (AV3), CCL 23 (HEp-2), CCL 25 (WISH), and CCL 62 (FL). The cells with no HeLa markers were (*a*) epithelial like—CCL 30 (RPMI 2650), CCL 185 (A549), CL 187 (LS 180), CL 188 (LS 174T), CCL 218 (WiDr), CCL 221 (DLD-1), CCL 227 (SW620), CCL 228 (SW480), CCL 229 (LoVo), CCL 233 (SW116), CCL 237 (SW948), and HTB 38 (HT-29); (*b*) lymphoblast like—CCL 86 (Raji), CCL 87 (Jijoye), CCL 114 (RPMI 7666), CCL 119 (CCRF-CEM), CCL 120 (CCRF-SB), CCL 155 (RPMI 8226), CCL 156 (RPMI 1788), CCL 159

(IM-9), CCL 213 (Daudi), CCL 214 (NC-37), CCL 240 (HL-60), CCL 243 (K-562), CCL 246 (KG-1), CCL 246.1 (KG-1a), and TIB 195 (CEM-CM3); (*c*) fibroblastic—CCL 75 (WI-38), CCL 127 (IMR-32), CCL 136 (RD), CCL 171 (MRC-5), CL 186 (IMR-90), and CCL 212 (MRC-9); and (*d*) other—CCL 220 (COLO 320DM), CCL 220.1 (COLO 320HSR), and HTB 64 (Malme-3M). Each cell line was cultured according to the procedures recommended by the American Type Culture Collection for that line (Hay et al. 1988). Two to 4 × 10^8 cells of each line were harvested and used for DNA extraction.

DNA Extraction and Electrophoresis

DNA was extracted according to standard procedures (Maniatis et al. 1982). Eight micrograms of DNA was digested with 50 units of *Hin*fI or *Hae*III in the reaction buffers supplied by the manufacturer (BRL, Bethesda, MD), in the presence of 4 mM spermidine trihydrochloride (Sigma Chemical) in a total reaction volume of 100 μl for 3 h at 37°C. After digestion, 10 μl of 0.2 M EDTA, 3 M sodium acetate (pH 5.2) was added, and the samples were extracted with an equal volume of buffer-saturated phenol and were ethanol precipitated. Pellets were washed in 70% ethanol, vacuum dried, and resuspended in 17 μl of gel running buffer (TAE; 40 mM Tris, 20 mM sodium acetate, 1 mM EDTA at pH 7.2). Samples were heated to 65°C for 15 min prior to loading. Electrophoresis was done in 1% agarose gel (20 cm × 20 cm) with a 20-slot, 1-mm-wide comb at 70 V in 1 × TAE running buffer until all fragments <1 kb had been electrophoresed off the gel.

Southern Blotting and Hybridization

Gels were soaked in 0.4 N NaOH for 30 min (Reed and Mann 1985) and then were blotted, by capillary action in the same buffer, onto Biotrace RP nylon membrane (Gelman Bioscience) for 8 h. After the blotting, the membranes were rinsed twice in 2 × SSC and baked for 2 h at 80°C under vacuum. Membranes were wetted in 0.1 × SSC for 10 min and then prehybridized in 0.5 M sodium phosphate at pH 7.2, 7% SDS, 1 mM EDTA, and 1% BSA (Church and Gilbert 1984) for 1–2 h at 65°C. Hybridization was carried out in the same buffer, with the addition of 1 × 10^6 cpm/ml of ^{32}P-labeled probe (see below), for 12–16 h at 65°C. Membranes were washed twice in 2 × SSC, 0.5% SDS at room temperature for 10 min and then twice in 0.1 × SSC, 0.5% SDS at 50°C for 30 min and finally were rinsed once in 2 × SSC at room temperature for 10 min. Membranes were kept moist and wrapped in Saran Wrap and were exposed to Kodak X-Omat AR film with one intensifying screen for 1–10 d.

The probe used was derived from the hypervariable probe 33.6 (Jeffreys et al. 1985*b*). The RF form of the M13-based 33.6 probe was digested with *Eco*RI and *Bam*HI, and the 600-bp fragment containing the 33.6 hypervariable sequence was subcloned into Bluescript vector pSK(−) (Stratagene). From this new construct, pSK33.6, the 600-bp *Eco*RI/*Bam*HI fragment containing the original 33.6 sequence was gel purified. Twenty-five nanograms of this isolated insert was rendered radioactive by random priming according to a method recommended by the manufacturer (Random Primer DNA Labeling Kit; Boehringer Mannheim).

For "high-resolution" gels, 4 μg of DNA prepared as described above in a 5 μl final volume was loaded on a 3.5%, 1.5-mm-thick, 30-cm-long polyacrylamide gel (19:1 acrylamide:bis) in 1 × TBE (0.89 M Tris-borate, 0.025 M EDTA) running buffer on a Hoeffer SE620 gel apparatus. Gels were run until the 80-bp marker was at the 25-cm point. Gels were blotted as above, but UV-Duralon (Stratagene) membrane was used. Filter was UV cross-linked by using a Stratalinker (Stratagene) at a total UV dose of 0.12 J. Membranes were hybridized and washed as described above. These gels allow separation, transfer, and resolution of DNA fragments in the 50–1,000-bp size range. Identical DNA samples were also run on a 2% agarose gel to clearly resolve fragments of 1,000–1,500 bp, so that DNA fragments in the overlapping molecular-weight range were detectable on both the agarose and acrylamide gels. Thus, we could be certain that we were not scoring bands which previously had been taken into account.

Gel Analysis

Each hybridized membrane was exposed to X-ray film for three different exposure times: 1–2 d to score high-copy bands, 3–5 d to score medium-copy bands, and 6–10 d to score low- or single-copy bands. To enable comparison of bands both within a gel and between gels, two molecular-weight standards were run on both sides of the gel. In addition, two cell lines (CCL 6 and CCL 243) were run on each gel. Comigration of DNA fragments of the same molecular mass was concluded when the difference in mobility of two fragments was ≤2 mm in two or more duplicate autoradiograms of 20-cm-wide gels loaded with 18 sample DNAs. This level of definition is affirmed by conformance of fragment migration in DNAs from the 10 HeLa derivatives (see figs. 1*B* and 1*D*). Different times of autoradiographic

Table 1

Human Cell Lines Typed with Minisatellite Probe 33.6 and *Hae*III Digestion

Cell Line			Date Deposited[b]	Passage Number[c]	Chromosome-number Range/Model Number[d]	No. of *Hae*III Fragments[e]	APD[f]	Frequency of *Hae*III DNA Fingerprint Best Estimate (upper bound)[g]
ATCC Number[a]	Name	Description						
Thirty-three distinct cell lines:								
CCL 2	HELA	Cervical carcinoma	11/61	102	70–164/82	18	75.1	1.8×10^{-17} (5.1×10^{-17})
CCL 30	RPMI 2650	Nasal septum tumor	1/63	22	43–92/46	16	74.2	1.8×10^{-16} (9.1×10^{-16})
CCL 75	WI-38	Diploid lung	6/67	10	33–93/46	19	74.6	7.0×10^{-17} (1.3×10^{-16})
CCL 86	RAJI	Burkitt lymphoma	3/67	106	43–192/46	18	77.6	4.1×10^{-18} (1.1×10^{-17})
CCL 87	JIYOYE	Burkitt lymphoma	3/67	57	42–92/46	13	76.6	2.4×10^{-16} (7.4×10^{-16})
CCL 114	RPMI 7666	Normal lymphoblast	4/68	UK	45–48/46	20	75.4	2.6×10^{-16} (2.6×10^{-16})
CCL 119	CCRF-CEM	Peripheral blood	12/68	UK	41–95/45	19	76.4	2.2×10^{-17} (4.0×10^{-17})
CCL 120	CCRF-SB	Peripheral blood	12/68	UK	42–99/46	18	78.5	1.2×10^{-18} (3.0×10^{-18})
CCL 127	IMR-32	Neuroblastoma	12/69	47	42–51/49	17	73.6	6.5×10^{-16} (2.7×10^{-15})
CCL 136	RD	Embryonic rhabdomyosarcoma	9/69	32	45–97/50	16	75.6	4.0×10^{-16} (2.0×10^{-15})
CCL 155	RPMI 8226	Myeloma	2/71	UK	57–135/68	17	77.5	3.8×10^{-17} (1.6×10^{-16})
CCL 156	RPMP 1788	Peripheral blood	11/72	UK	43–89/45	17	75.6	4.8×10^{-17} (2.0×10^{-16})
CCL 159	IM-9	Lymphoblast	10/73	UK	43–48/46	17	80.0	1.0×10^{-17} (3.1×10^{-17})
CCL 171	MRC-5	Diploid lung	6/72	14	44–92/46	17	80.2	8.3×10^{-18} (2.3×10^{-17})
CCL 185	A549	Lung cancer	7/76	73	44–102/65	18	75.3	1.6×10^{-16} (5.2×10^{-16})
CCL 186	IMR-90	Diploid lung	9/76	7	34–48/46	17	81.7	6.6×10^{-19} (2.6×10^{-18})
CL 187	LS 180	Colon adenocarcinoma	3/77	34	42–47/45	21	71.6	8.8×10^{-16} (8.8×10^{-16})
CCL 212	MRC-9	Lung	3/79	2	40–47/46	19	75.3	1.1×10^{-16} (2.1×10^{-16})
CCL 213	Daudi	Burkitt lymphoma	10/78	UK	43–47/46	18	84.2	2.4×10^{-21} (6.6×10^{-21})
CCL 214	NC-37	Lymphoblast	12/78	UK	40–51/48	17	79.3	5.8×10^{-18} (2.2×10^{-17})
CCL 218	WIDR	Colon adenocarcinoma	4/79	19	38–84/72	17	71.8	4.6×10^{-15} (1.8×10^{-14})
CCL 220	COLO 320DM	Colon adenocarinoma	9/79	7	41–63/53	13	77.4	2.2×10^{-17} (8.5×10^{-17})
CCL 221	DLD-1	Colon adenocarcinoma	7/79	18	40–51/46	18	73.4	4.0×10^{-16} (1.3×10^{-15})
CCL 227	SW620	Metastasized adenocarcinoma	11/78	83	45–53/50	20	76.0	2.7×10^{-18} (2.7×10^{-18})
CCL 229	LOVO	Colon adenocarcinoma	1/79	25	41–54/49	18	70.3	6.6×10^{-15} (1.8×10^{-14})
CCL 233	SW1116	Colon adenocarcinoma	11/78	37	50–62/60	18	78.5	1.8×10^{-17} (5.6×10^{-17})
CCL 237	SW948	Colon adenocarcinoma	11/78	51	47–68/67	16	79.1	2.1×10^{-17} (1.2×10^{-16})
CCL 240	HL-60	Promyelocytic leukemia	3/82	UK	43–48/46	17	74.9	5.3×10^{-16} (2.6×10^{-15})
CCL 243	K-562	Chronic myelogenous leukemia	3/83	UK	64–77/69	15	82.6	1.7×10^{-19} (7.0×10^{-19})
CCL 246	KG-1	Acute myelogenous leukemia	5/84	UK	44–49/47	20	77.5	2.9×10^{-17} (2.9×10^{-17})
HTB 38	HT-29	Colon adenocarcinoma	1/82	125	68–72/71	18	74.4	5.4×10^{-16} (1.6×10^{-15})
HTB 64	Malme-3M	Metastasized melanoma	1/82	22	NT	16	83.1	1.3×10^{-19} (4.9×10^{-19})
TIB 195	CEM-CM3	Acute lymphoblastic leukemia	11/83	UK	NT	16	79.6	1.2×10^{-17} (3.4×10^{-17})
Nine HeLa derivatives related to CCL 2:								
CCL 5	L-132	Embryonic lung	3/62	15	54–141/71	17	2.86	1.3×10^{-17} (4.7×10^{-17})
CCL 6	Intest 407	Embryonic intestine	10/62	273	55–153/76	18	.32	1.8×10^{-17} (5.1×10^{-17})
CCL 13	Chang LIV	Liver	12/62	255	62–82/70	18	.32	1.8×10^{-17} (5.1×10^{-17})
CCL 17	KB	Oral carcinoma	6/63	361	63–150/74	18	.32	1.8×10^{-17} (5.1×10^{-17})
CCL 18	Detroit 98	Sternal marrow	9/63	117	38–100/63	18	.32	1.8×10^{-17} (5.1×10^{-17})
CCL 21	AV3	Amnion	9/63	270	58–95/74	18	.32	1.8×10^{-17} (5.1×10^{-17})
CCL 23	HEP-2	Larynx carcinoma	12/63	350	59–195/76	18	.32	1.8×10^{-17} (5.1×10^{-17})
CCL 25	WISH	Amnion	6/63	167	56–77/64	18	.32	1.8×10^{-17} (5.1×10^{-17})
CCl 62	FL	Amnion	1/66	480	68–80/71	18	.32	1.8×10^{-17} (5.1×10^{-17})

Table 1 (continued)

ATCC Number[a]	Cell Line: Name	Cell Line: Description	Date Deposited[b]	Passage Number[c]	Chromosome-number Range/Model Number[d]	No. of *Hae*III Fragments[e]	APD[f]	Frequency of *Hae*III DNA Fingerprint Best Estimate (upper bound)[g]
Four duplicate clones:[h]								
CL 188	LS 174t	Colon adenocarcinoma	3/77	104	36–46/45	21	4.76	1.5×10^{-15} (1.5×10^{-15})
CCL 220.1	COLO 320HS	Colon adenocarcinoma	9/79	7	49–61/53	13	.00	2.2×10^{-17} (8.5×10^{-17})
CCL 228	SW480	Colon adenocarcinoma	11/78	96	51–111/55	20	.00	2.7×10^{-18} (2.7×10^{-18})
CCL 246.1	KG-1A	Acute myelogenous leukemia	5/84	UK	44–49/47	20	.00	2.9×10^{-17} (2.9×10^{-17})

[a] From American Type Culture Collection *Cagalogue of Cell Lines and Hybridomas* (Hay et al. 1988), in which a more detailed description of each line can be found (CCL = Certified Cell Line; CL = Cell Line, patent deposit; HTB = Human Tumor Cell Bank; TIB = Tumor Immunology Bank).

[b] When culture received at ATCC.

[c] At deposition. UK = unknown.

[d] Actual karyotype analysis of each line, as derived from 100–200 metaphases, reported as range over modal number. NT = not tested.

[e] Total number of bands resolved for *Hae*III

[f] Computed by averaging the PD of a given cell line over various pairings: for the 33 distinct cell lines the average is over 32 possible choices for pairing; for the nine HeLa derivatives, the average is over the nine choices for pairing, including CCL 2; and for the four duplicate clones, the PD against the duplicate is shown.

[g] As discussed in text for *Hae*III only.

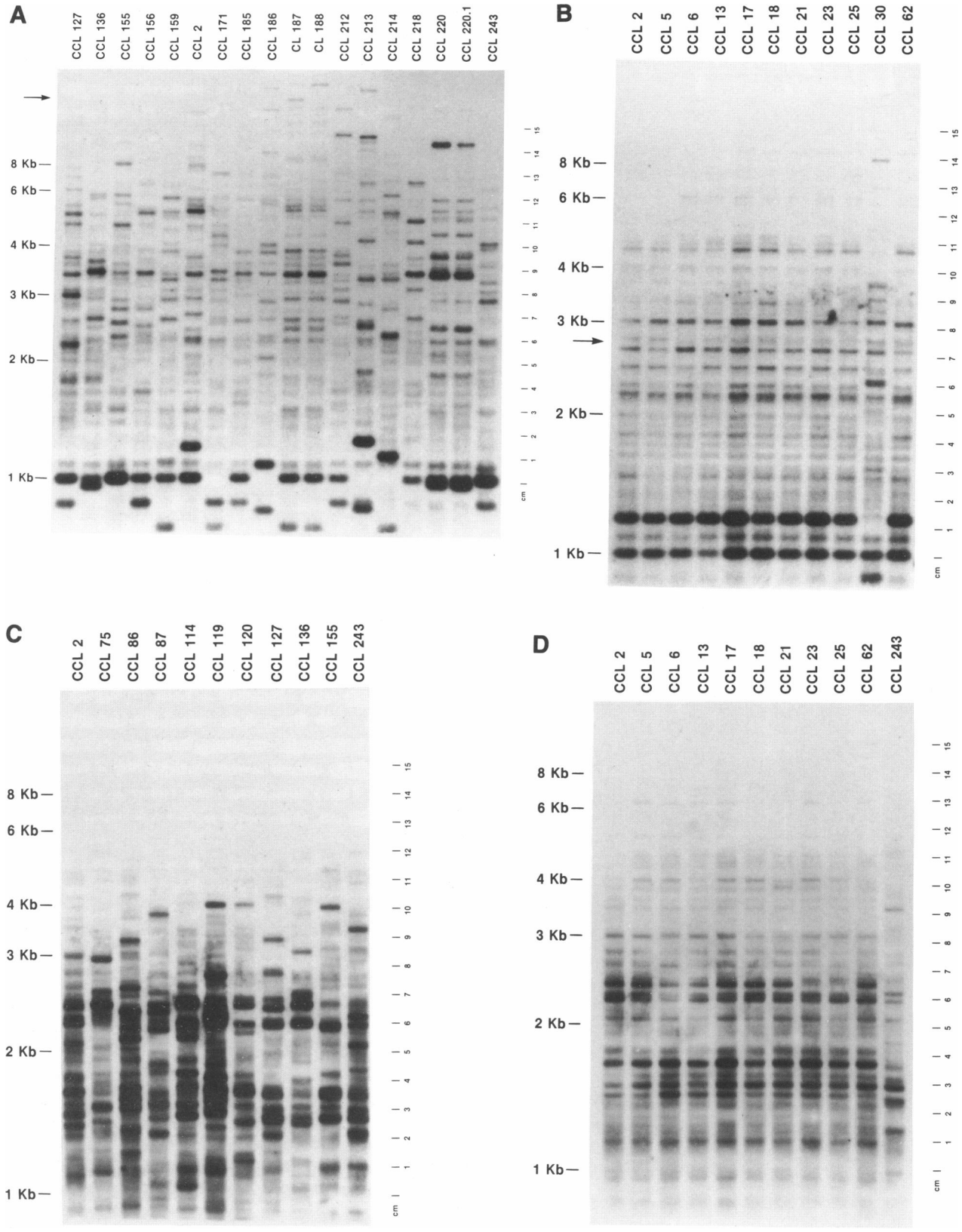
[h] Duplicates for clones CL 187, CCL 220, CCL 227, and CCL 246, respectively.

development (1–10 d) were used for diagnostic gels, and only fragments visible on short (1–2-d) exposures were included in the analysis (Lander 1989). A total of 121 distinct fragments within the molecular-weight range of 1.0–15 kb were scored. The composite phenotype of each DNA fragment of an individual comprised the DNA fingerprint.

Genetic Model and Probability Calculations

The genetic calculations in the present paper are based on the following genetic model: It is assumed that the alleles tracked by the hypervariable probes are in Hardy-Weinberg equilibrium and that fragments (bands) of the same molecular weight and intensity represent the same allele of a particular locus, although it is recognized that this may not always be true (Lander 1989). Each DNA fragment corresponds to a unique codominant allele on one of L loci, but we do not know which bands correspond to which loci. Nonetheless, the observed DNA fragment frequency, F_i, is related to the corresponding population allele frequency P_{lj}, by $F_i = P_{lj}^2 + 2P_{lj}(1-P_{lj})$. Here P_{lj} is the population frequency, of allele j on locus l, that corresponds to band i, $i = 1,2 \ldots 121$. Solving, we estimate that $P_{lj} = 1 - (1-F_i)^{1/2}$ on the basis of band frequency data. Note that $\Sigma_l\Sigma_j P_{lj} = L$, so that we can estimate the total number of loci as $L = 121 - \Sigma_{i=1}^{121}(1-F_i)^{1/2}$, which in our data is 9.406. On the basis of this calculation, and because some persons were observed to have 20 bands, we assume that $L = 10$. If our data included fragments which are comigrating bands stemming from two different loci, our calculation would underestimate the actual number of loci. This calculation does not require knowing which alleles are associated with each locus (see Appendix).

If we knew which alleles were associated with each locus, the probability of a given DNA fingerprint ("composite band pattern") would be the product, over all loci, of the probability of the locus-specific genotypes. If A_l alleles are associated with locus l, there are $C(A_l + 1, 2)$ distinguishable band patterns associated with that locus when the notation $C(a,b) = a!/[b!(a-b)!]$ is used. The probability that the locus-specific band pattern will have bands that correspond to alleles j and k on locus l is given by P_{lj}^2 if $j = k$ and by $2P_{lj}P_{lk}$ if $j \neq k$. There are $\Pi_{l=1}^{10} C(A_l + 1, 2)$ composite band patterns, made up of all possible combinations of locus-specific band patterns. If it is assumed that allele frequencies at various loci are independent (see below), the probability of a composite band pattern is the prod-


A
CCL 127
CCL 136
CCL 155
CCL 156
CCL 159
CCL 2
CCL 171
CCL 185
CCL 186
CL 187
CL 188
CCL 212
CCL 213
CCL 214
CCL 218
CCL 220
CCL 220.1
CCL 243
8 Kb
6 Kb
4 Kb
3 Kb
2 Kb
1 Kb
B
CCL 2
CCL 5
CCL 6
CCL 13
CCL 17
CCL 18
CCL 21
CCL 23
CCL 25
CCL 30
CCL 62
8 Kb
6 Kb
4 Kb
3 Kb
2 Kb
1 Kb
cm
C
CCL 2
CCL 75
CCL 86
CCL 87
CCL 114
CCL 119
CCL 120
CCL 127
CCL 136
CCL 155
CCL 243
8 Kb
6 Kb
4 Kb
3 Kb
2 Kb
1 Kb
cm
D
CCL 2
CCL 5
CCL 6
CCL 13
CCL 17
CCL 18
CCL 21
CCL 23
CCL 25
CCL 62
CCL 243
8 Kb
6 Kb
4 Kb
3 Kb
2 Kb
1 Kb
cm

uct, over loci, of the probabilities of the locus-specific band patterns.

Because we do not know exactly which bands are associated with which loci, a somewhat involved averaging process, described in the Appendix, is used to estimate the probability of a given composite band pattern in the general population. In addition, we compute an upper bound for that probability, as outlined in the Appendix. We assume throughout that our 33 distinct cell lines are representative of all cell lines, even though no random sampling plan was used to obtain the cell lines.

Methods for estimating the probability of two or more common composite band patterns among N randomly selected individuals are also given in the Appendix. The genetic models were executed with results of *Hae*III. Similar evaluations were performed for *Hin*fI, under a model with nine loci, as $L = 114 - \sum_{i=1}^{114} (1-F_i)^{1/2} = 8.342$ loci.

Results

The DNA fingerprints of 46 human cell lines listed in table 1 were determined using the minisatellite probe designated 33.6 (Jeffreys et al. 1985*b*). The selected cell lines included (1) nine cell lines (CCL 2, CCL 5, CCL 6, CCL 13, CCL 17, CCL 18, CCL 21, CCL 23, CCL 25, and CCL 62) previously shown, on the basis of chromosome markers or allozyme genetic signature, to be derivatives of HeLa; (2) four pairs of cell lines (CL 187/188, CCL 220/220.1, CCL 227/228, and CCL 246/246.1) derived from the same individual at different times; and 28 cell lines of different origins considered, on the basis of their history, their allozyme genetic signature, or other specific cell characteristics, to be unique. In all, there were 33 different individuals represented in the survey.

Genomic DNA from each cell line was digested with *Hae*III and *Hin*fI, separated electrophoretically, transferred to nylon filters, and hybridized with the radiolabeled hypervariable probe, 33.6. Representative autoradiograms for cell lines are illustrated in figure 1. Relevant frequency and population genetic estimates are summarized in table 2. The total number of fragments resolved with both enzymes for each cell line varied from 28 to 37, but the average was 34 bands/individual DNA sample. As expected from the structure of minisatellites (Jeffreys et al. 1985*b*), the variation of fragment mobility between different individuals was great. Among the 33 individual lines, we could resolve a total of 121 *Hae*III fragments and 114 *Hin*fI fragments over a molecular-weight range of 1–15 kb.

The frequency of appearance of each of the unique fragments in the sample cell lines varied from .03 to .78, with the majority (65%) having an incidence of $\leq$.12 (fig. 2). The average band frequency for both enzymes was similar: .144 (range .03–.69) for *Hae*III and .137 (range .03–.78) for *Hin*fI. None of the fragments was present in all individuals.

In order to compare individuals directly, the DNA fingerprint of each cell line was compared with that of every other cell line. The extent of quantitative genetic difference was calculated as the percent difference (PD) in resolved DNA fragments, i.e., the number of fragments which were different between two cells divided by the total number of fragments present in both cell lines multiplied by 100. For each cell listed in table 1, we computed the PD of *Hae*III and *Hin*fI fragments between that line and each of the other cell lines. We present (1) the PD of *Hae*III fragments in DNA fingerprints between the unique cell lines (table 3) and (2) the same estimates between pairs chosen from 10 HeLa cell derivatives and from four pairs of unique cell lines where each pair was derived from the same patient (table 4). Comparison of the PD values of these two tables shows the extreme power of the DNA fingerprint in implicating identity of individual cell lines. PDs among cells derived from the same individuals are very small (0%–3%), while those between unrelated cell lines are very high (47%–100%), with no overlap in 528 pairwise comparisons. The genetic uniqueness of each cell's DNA fingerprint is apparent by computation of the APD of each cell line versus every other, unrelated cell line (see table 1).

The four pairs of cell lines, which included separate samples from the same individual, had a DNA finger-

Figure 1 Autoradiograms of DNA fingerprint of cell lines by using probe 33.6. These are 3–4-d exposures (see Material and Methods). *A*, *Hae*III-digested DNA. Cell lines 187 and 188 were different explants from the same patient. Similarly, CCL 220 and 220.1 were derived from the same donor at different biopsies. The arrow indicates a single novel fragment present in CCL 188 but absent in CCL 187; otherwise all bands were shared in the two respective pairs. *B*, *Hae*III-digested DNA of HeLa (CCL 2) and HeLa derivatives (CCL 5, CCL 6, CCL 13, CCL 17, CCL 18, CCL 21, CCL 23, CCL 25, and CCL 62). The arrow indicates the HeLa fragment present in all HeLa derivatives except CCL 6 (see text). *C*, *Hin*fI-digested DNA of 10 representative unrelated cell lines. *D*, *Hin*fI-digested DNA of HeLa and HeLa derivatives.

Table 2

Observed Population Genetic Parameters in Population of 33 Individual Human Cell Lines when Minisatellite Probe 33.6 Is Used

	Restriction Enzyme	
	*Hae*III	*Hin*fI
Total no. of unique fragments	121	114
Mean ± SD no. of fragments resolved per individual	17.4 ± 1.8	15.6 ± 2.0
Mean frequency of fragment in population: probability that fragment in A is present in B	.144	.137
Allele frequency:		
Median	.047	.063
Range	.015–.450	.015–.539
Estimated DNA fingerprint frequency in human population:		
Median	2.9×10^{-17}	5.6×10^{-17}
Range	2.4×10^{-21} to 6.6×10^{-15}	1.2×10^{-19} to 1.2×10^{-14}
Average heterozygosity:		
Model I (see text)	.85	.87
Model II (see text)	.84	.87
APD among unrelated individuals	76.9	81.1

print virtually identical to each other but a highly distinctive phenotype from other cell lines (figs. 1*A* and 1*C* and table 4). The nine HeLa derivatives had almost identical fingerprints to the HeLa standard, with only 1 of the 36 resolved bands differing (APD = 0.6%) (see figs. 1*A* and 1*C*). The reason for these differences is unknown but could represent chromosome loss or rearrangement, which is known to occur in the culturing of human tumor cells (Lavappa et al. 1976).

The resolved DNA fragments have been shown to represent multiple alleles of several human loci dispersed on numerous human chromosomes (Jeffreys et al. 1985*c*). The fragment size variation detected by this probe reflects allelic differences in the number of tandem repeats of a short consensus DNA sequence present in each allele/locus. If we presume that the alleles at minisatellite loci assort independently and according to a multinomial distribution (i.e., conform to the Hardy-Weinberg equilibrium) in human populations, then the estimated allelic and genotypic frequencies can be used to predict genotypic incidence in a population of human cells. We tested for evidence of allelic independence (or association) by computing χ^2 values for each of 7,260 pairwise combinations of 121 fragments in the population, by using a method designed to account for large numbers of comparisons (Schweder and Spjotvoll 1982). This test does not assume that a particular band corresponds to a particular allele at a particular locus, but it is designed to detect nonrandom association of any two fragments. Nonrandom association would occur if two fragments were in linkage disequilibrium or if the restriction endonuclease cut within a minisatellite allele and produced adjacent flanks (Wong et al. 1986). No evidence of fragment association was detected with tests using all 121 alleles or when only abundant alleles ($F \geq .18$) were tested for association, a finding supporting our assumption of independent allelic assortment in the cell populations. We therefore can estimate the allele frequency of any fragment, p, as $p = 1 - (1 - F)^{1/2}$, where F is the corresponding fragment frequency. The median value of p was .047 (range .015–.450) for *Hae*III and .063 (range .015–.539) for *Hin*fI (table 2).

An approximation of the individual frequency of a given DNA fingerprint is derived from the cumulative genotype frequencies (both homozygotes and heterozygotes) derived from the genetic model described in Material and Methods. In this model for *Hae*III, we assume that multiple alleles at 10 polymorphic loci segregate independently in the population. Nine loci are assumed for *Hin*fI. In the special case of heterozygosity at each of these loci, the frequency would be estimated simply by $\Pi_i^L 2pq$, where L is the number of loci and p and q are alternative allele frequencies at the ith locus (as in Jeffreys et al. 1985*c*). Since individuals show variation in the number of resolved fragments, the present model allows the possibility that some of these fragments represent homozygous alleles. We used a weighted averaging process based on allele frequencies (see Appendix) to account for the effects of homozygosity on

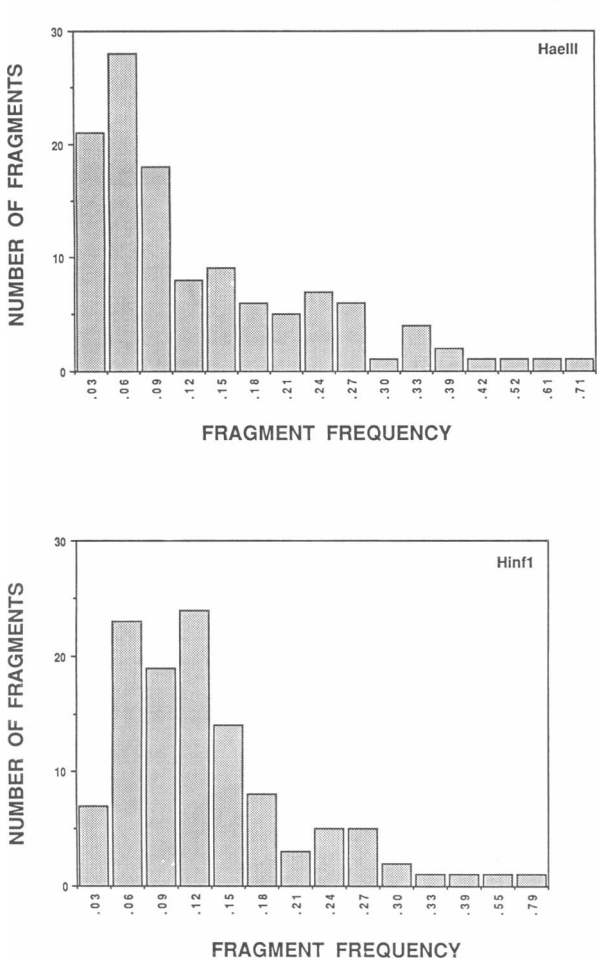


Figure 2 Frequency of occurrence of different DNA fragments among the 33 distinct individual cell lines. The frequency was determined by dividing the number of cell lines exhibiting a particular band by the total number of unique cell lines tested (33).

fingerprint frequencies. We present the derived fingerprint frequency for each cell line in table 1. The median DNA fingerprint frequency was 2.9×10^{-17} for *Hae*III (range 2.4×10^{-21} to 6.6×10^{-15}) and 5.6×10^{-17} for *Hin*fI (range 1.2×10^{-19} to 1.2×10^{-14}). These estimates represent the probability that a particular DNA fingerprint will be matched in a second, unrelated cell line. We believe that these probabilities are conservative, because they usually exceed similar estimates based on the more realistic model of $N = 11$ loci, where certain alleles are invisible because they are below the resolvable fragment size ($\leq$1.5 kb) of the gel system (see below and Appendix).

In addition to our best estimate, based on averaging, we present an upper bound on the probability of each DNA fingerprint (table 1), as calculated in the Appendix. Note that the upper bound tends to be about three times larger than the estimated probability, but even the upper bound is tiny.

Average heterozygosity over all loci depends on how those few alleles that occur in relatively high ($\leq$.20) frequency are distributed over loci. Because their distribution is unknown, we provide two estimates based on two extreme models (see Appendix). Model I assumes that abundant alleles are clustered to a minimum number of loci; model II assumes that abundant alleles are distributed equally over all loci. Using the procedure described in the Appendix for the Hae III data, model I provides an estimate of average heterozygosity of .85, and for model II average heterozygosity is .84. For *Hin*fI, this estimate of average heterozygosity is .87 under both models.

The probability of a chance match of DNA fingerprints in groups of N randomly selected individuals has been computed using a modification of the "generalized birthday problem" methodology previously applied to allozyme genetic signatures (Gail et al. 1979; O'Brien et al. 1980). Unlike theory for the classical "birthday problem," where each birthday is equally likely (i.e., 1/365), the present methods account for the appreciable variation in frequencies of individual DNA fingerprints in human populations (see table 1 and Appendix). As before, two extreme models relating to the dispersal of abundant alleles were considered. Model I assumes that abundant alleles are clustered to a minimal number of loci and produces more homogeneous fingerprint frequencies; model II assumes equal distribution of more abundant alleles among all loci and would produce relatively heterogeneous DNA fingerprint frequencies. The results (table 5) are quite dramatic even when thousands of individuals are typed. The two models gave very similar estimates, indicating that the dispersal of abundant alleles is of small consequence in the computation. Thus, if 100,000 randomly selected cells were typed, the probability of match by chance remains vanishingly small ($3–5.7 \times 10^{-6}$).

Since DNA fingerprinting gels are routinely electrophoresed until all fragments <1 kb are run off the gel, it is not known what effect those missing fragments would have on our calculations and genetic modeling. In order to address this point, we performed "high-resolution" DNA fingerprinting analysis on DNA from the 33 unique cell lines (see Material and Methods). Polyacrylamide-gel electrophoresis was used to resolve all detectable *Hae*III fragments in a 50–1,000-bp range. The average number of additional fragments detected

Table 3

PD[a] in *HaeIII* Fragments Retained between 33 Cell Lines Derived from Unrelated Donors

	CCL																CL	CCL													HTB	
	2	30	75	86	87	114	119	120	127	136	155	156	159	171	185	186	187	212	213	214	218	220	221	227	229	233	237	240	243	246	38	64
CCL:																																
30	66																															
75	74	66																														
86	84	71	73																													
87	81	66	69	61																												
114 ...	85	72	64	68	82																											
119 ...	63	71	74	84	69	74																										
120 ...	68	77	84	78	74	74	62																									
127 ...	67	70	67	85	60	68	72	77																								
136 ...	77	63	66	77	93	67	71	82	70																							
155 ...	78	58	67	82	73	62	67	77	82	64																						
156 ...	72	76	61	89	73	84	72	83	53	82	77																					
159 ...	89	88	78	77	85	68	83	83	88	76	77	82																				
171 ...	78	82	83	77	80	73	78	66	100	76	77	94	53																			
185 ...	78	53	68	72	68	84	85	78	71	88	83	66	83	77																		
CL 186 ..	78	76	83	77	87	84	72	77	88	82	82	82	82	82	83																	
CCL:																																
187 ...	60	68	70	80	71	76	60	74	68	62	84	68	79	68	69	68																
212 ...	79	83	84	73	88	74	74	84	72	71	78	72	72	72	68	78	70															
213 ...	73	88	84	78	94	90	84	89	89	88	89	77	77	83	84	83	90	84														
214 ...	89	70	72	54	73	78	89	83	88	82	77	88	71	77	71	94	79	78	89													
218 ...	61	70	67	66	73	68	78	71	53	58	88	65	65	82	71	82	63	67	83	82												
220 ...	75	66	75	87	77	70	75	74	80	72	80	80	80	87	74	80	65	81	87	80	73											
221 ...	84	77	78	72	81	53	89	83	66	77	77	77	77	83	72	77	80	62	89	71	60	68										
227 ...	74	72	69	84	70	80	80	84	68	67	84	68	84	85	79	100	76	74	79	78	68	70	68									
229 ...	78	77	84	72	81	84	78	89	77	82	89	77	83	89	67	77	80	68	78	83	83	68	78	74								
233 ...	57	71	73	78	48	79	68	72	60	77	77	60	94	83	78	77	59	73	67	71	71	61	67	53	67							
237 ...	77	88	77	77	79	83	94	88	76	69	94	76	88	76	71	88	68	66	82	76	76	86	77	72	82	65						
240 ...	78	70	67	83	90	89	83	71	53	76	71	65	71	88	54	88	68	78	94	71	71	80	77	73	71	71	82					
243 ...	88	94	82	94	93	89	82	88	88	87	75	81	75	81	82	81	78	82	82	88	88	93	70	71	88	82	87	69				
246 ...	59	89	80	84	82	75	90	84	68	94	78	78	95	89	74	84	66	80	79	73	73	94	68	85	74	63	61	73	77			
HTB:																																
38	68	77	73	78	81	74	84	78	60	77	83	66	89	89	72	89	74	73	83	49	49	81	56	63	78	67	71	77	82	47		
64	89	81	89	88	94	72	94	82	82	69	70	88	88	82	88	88	89	88	88	94	94	79	59	83	82	82	81	76	74	78	64	
TIB 95 ..	77	88	94	82	86	72	49	77	94	81	82	88	82	70	88	64	62	66	77	82	72	79	77	89	77	71	88	100	74	78	88	81

[a] Mean PD for 528 comparisons of 33 cell lines = 76.87; median PD = 77.14; PD range = 43.37–100.0; SD = 9.46. Compare with Note to table 4.

Table 4

PD[a] in *HaeIII* Fragments Retained between Multiple Cell Lines Derived from Five Unrelated Individuals

	CCL										CL		CCL				
	2	5	6	13	17	18	21	23	25	62	187	188	220	220.1	227	228	246
CCL:																	
5	0																
6	3	3															
13	0	0	3														
17	0	0	3	0													
18	0	0	3	0	0												
21	0	0	3	0	0	0											
23	0	0	3	0	0	0	0										
25	0	0	3	0	0	0	0	0									
62	0	0	3	0	0	0	0	0	0								
CL:																	
187 (LS180)	68	68	73	68	68	68	68	68	68	68							
188 (LS174t)	68	68	73	68	68	68	68	68	68	68	3						
CCL:																	
220 (COLO 320 DM)	83	83	83	83	83	83	83	83	83	83	79	79					
220.1 (COLO 320 HSR)	83	83	83	83	83	83	83	83	83	83	79	79	0				
227 (SW620)	73	73	72	73	73	73	73	73	73	73	74	74	68	68			
228 (SW480)	73	73	72	73	73	73	73	73	73	73	74	74	68	68	0		
246 (KG-1)	74	74	80	74	74	74	74	74	74	74	64	63	68	68	69	69	
246.a (KG-1A)	74	74	80	74	74	74	74	74	74	74	64	64	68	68	69	69	0

[a] Mean PD for HeLa derivatives = .57, median PD = .0; PD range = .00–2.86; SD = 1.16. For the four pairs of cell lines derived from the same patient, the inter se mean PD = 1.19, median PD = .00; PD range = .00–4.76; SD = 2.38. Compare with Note to table 3.

in this low-molecular-weight range was 4.4 (range 3–6). We resolved 11 new unique bands, with an average band frequency of .25. Overall, the additional 11 bands represent 9% of the total fragments scored. The number of additional loci represented by these low-molecular-weight fragments can be estimated as $L = 11 - \Sigma_{i=1}^{11}$ $(1-F_i)^{1/2} = 1.7$. The sum of fragments resolved by conventional and "high-resolution" DNA fingerprints was considered in extended calculations of fingerprint probabilities discussed in the Appendix. Inclusion of these additional fragments and predicted additional loci do not substantially affect the derived probability values (compare table A1 with tables 2 and 5). Thus, calculations presented above that are based on scoring of agarose-gel fragments of ≥1 kb provide an accurate sampling of the hypervariable loci represented in the human genome.

Table 5

Estimated Probability of Chance Identity of Two DNA Fingerprints in a Population of *N* Individuals

N	Model I	Model II
2	$<10^{-7}$	$<10^{-7}$
10	$<10^{-7}$	$<10^{-7}$
10^2	$<10^{-7}$	$<10^{-7}$
10^3	$<10^{-7}$	$<10^{-7}$
10^4	$<10^{-7}$	$<10^{-7}$
10^5	3.0×10^{-6}	5.7×10^{-6}
10^6	3.0×10^{-5}	5.6×10^{-4}
10^7	3.0×10^{-3}	5.5×10^{-2}

NOTE.—Model I assumes clustering of abundant alleles to a minimum number of loci producing more homogeneous DNA fingerprint frequencies; model II assumes equal distribution of the more abundant alleles (i.e., $P \geq .20$) among all loci and produces relatively heterogeneous composite frequencies.

Discussion

A survey of 33 unique cell lines demonstrated that DNA fingerprinting using human minisatellite probes is a verifiable and rigorous method for discriminating individuals and for detecting genetic identity among human cell lines. On the basis of band frequencies in the tested cells, we detect an average allele frequency of .07 and a median composite DNA fingerprint frequency of 2.9×10^{-17}. This frequency approximates the probability that a randomly selected cell line would match a given DNA fingerprint by chance.

In order to quantitate the extent of genetic difference between individuals, we computed the PD, in fragment retention, between each individual cell line in the survey. The APD between unrelated cells was 76.9% (range 70.3%–84.2%). For three pairs of cell lines derived from the same patient, the DNA fingerprints were identical (PD = .0). In a comparison of nine cell lines shown previously to be HeLa derivatives, eight had a DNA fingerprint which was identical to that of prototype HeLa (CCL 2), and one (CCL 6) had a single fragment difference. Each of these HeLa derivatives had an APD of 75.1% (79.5% for CCL 6) versus all other cells. The loss of a common fragment in the HeLa derivative CCL 6 is not easy to explain but could reflect derived monosomy in certain cells (those which lack the fragment) of chromosome segments that contained the missing allele. This explanation would be supported by the dynamic nature of heteroploid human tumor cells in culture, particularly in HeLa (Lavappa et al. 1976; Nelson-Rees 1980). Nonetheless, the extreme similarity (PD = 97%–100%) of these derivative cell lines is dramatically different from the extent of band sharing between unrelated cells (~20%). Among first-order relatives (parent-offspring and siblings) one would expect only 60% identity, so a PD of 0%–3% would most certainly be interpreted as identity plus cell culture-associated allelic loss rather than as consanguinity in human populations.

We have used a simplified genetic model with 10 loci to account for findings and to carry out probabilistic calculations. The assumption of no linkage disequilibrium is crucial to our calculations, but our data, which have limited power to detect disequilibrium, do not indicate its presence. The fact that we do not know which alleles reside on which loci does not prevent our estimating individual allele frequencies—or even our getting reasonable estimates of the probabilities of any particular fingerprint. The present method is more realistic than a similar estimation of DNA fingerprint frequencies in human populations that was developed by Jeffreys et al. (1985*b*, 1985*c*), because the present procedure accounts for the occasional presence of homozygosity that will occur in screens of VNTR loci with multilocus profiles.

Recently, on the basis of data on single-locus probes, the question has been raised (Lander 1989) as to whether the VNTR loci are in Hardy-Weinberg equilibrium. In the multilocus system used here it is impossible to calculate the actual frequency of heterozygotes and homozygotes detected by the hypervariable sequence probes. We have assumed that the alleles detected are in Hardy-Weinberg equilibrium, since none of the tests applied (see above) have indicated any significant fragment associations. However, if more genetic data on which alleles belonged at which loci were available, we could obtain improved estimates of allele frequencies and the probabilities of individual fingerprints. Lacking this information, we have taken a "worst case" and "best case" approach in table 2, in an effort to bracket the true probabilities (see Appendix).

A striking aspect of the HeLa DNA fingerprints is their relative stability over prolonged culturing in a variety of conditions. Each of the HeLa derivative cell lines has been maintained as a separate culture for >20 years and often through hundreds of passages (see table 1). Even with marked variation in cytogenetic, biochemical, and differentiated functions (Nelson-Rees et al. 1980), these lines give a DNA fingerprint that is virtually identical to the HeLa standard (figs. 1*B* and 1*D* and table 4). Thus, it appears that these cell lines were contaminated with HeLa early in their history and that the cultures subsequently deposited with the American Type Culture Collection for distribution were simply HeLa lines. In addition, the four pairs of cell lines obtained from the same patients at different times and under different conditions showed nearly identical fingerprints. Thus, a DNA fingerprint remains heritable and stable even after a heteroploid transformation and continuous in vitro passage. It appears, therefore, that DNA fingerprinting is a reliable method for cell-line individualization, a method which exceeds the genetic resolution of other techniques previously employed in cell-culture monitoring. The application of this procedure represents an extremely accurate and relatively straightforward method to assure the continued integrity of culture collections and to ensure that future cell-line cross contamination or misidentification can be quickly recognized.

Acknowledgments

We are grateful to Drs. William Modi and Scott Baken for critical comments on early versions of the manuscript. We thank Drs. Robert Wayne and Niles Lehman for helpful discussions, and we thank Dr. Alec Jeffreys for providing the human minisatellite probe 33.6. This project has been funded at least in part with Federal funds from the Department of Health and Human Services under contract NO1-CO-74102 with Program Resources, Inc. The content of this publication does not reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Appendix

Estimates of Frequencies for Individual Fingerprints and for the Chance of Common Fingerprints among Groups of Randomly Selected Individuals

Individual Fingerprint Probabilities Based on the Simple Model with 10 Loci

As in the text, we let F_i be the observed frequency of band i, and we define the estimated allele frequency $p_i = 1 - (1 - F_i)^{1/2}$ corresponding to band i. If we knew that this allele was the jth allele on locus l, we would identify $p_i - P_{lj}$ in the earlier notation of the text.

First, consider a pattern with exactly 20 bands. Then each of the 10 loci must be heterozygous. If we knew which bands were associated with which loci, we would estimate the probability of that composite band pattern as $\Pi_{l=1}^{10} 2P_{lj}P_{lk} = 2^{10}\Pi p_i$, where the last product is over the bands present. Note that this estimate does not require knowing which bands are associated with which loci. Suppose instead that only 19 bands are present, and suppose that the homozygous band is labeled i. Then the probability of the composite band pattern is estimated by $p_i^2 2^9 \Pi_{j \neq i} p_j = p_i 2^9 \Pi p_j$, where the product is over all distinct bands (19 in this case). In the absence of knowledge about which band is homozygous, our best estimate of $p_i 2^9 \Pi p_j$ is obtained by replacing p_i by its expected value. The expected value of p_i over all possible choices for the unknown homozygous allele, when one of the observed bands must represent an homozygous allele, is $\Sigma p_i^2 / \Sigma p_i$, where sums are over all observed bands. In general, if B bands are observed, we define $D = 20 - B$ and estimate the probability of the composite band pattern by $K 2^{10-D} \Pi p$ where the product is over all observed bands, where $K = \Sigma p_{i_1}^2 p_{i_2}^2 \ldots p_{i_D}^2 / \Sigma p_{i_1} p_{i_2} \ldots p_{i_D}$, and where the sums are over all $C(B,D)$ ways of selecting D homozygous alleles from B alleles. Of course, if we knew which bands were associated with which loci, we would not need to go through this averaging process, since we would know exactly which bands represented homozygous alleles.

We obtain an upper bound on the probability of the observed composite band pattern as follows. For 19 bands, the upper bound is (largest p_j)$2^9 \Pi p_j$ where (largest p_j) is the largest p_j among the 19 observed bands and where the product is over all observed bands. For 18 observed bands the upper bound is (largest p_j) (next largest p_j)$2^8 \Pi p_j$. For fewer bands, the upper bound is calculated similarly.

Estimates of Probability of Two or More Common DNA Fingerprints among N Randomly Selected Individuals, Based on the Simple Model with 10 Loci

Suppose we know that there are A_l alleles on locus l for $l = 1.2, \ldots 10$. Then there are $s = \Pi_{l=1}^{10} C(A_l + 1,2)$ possible DNA fingerprints (composite band patterns). Let γ_i be the frequency of fingerprint i, for $i = 1,2, \ldots s$. Gail et al. (1979) give a general method to compute the probability of no common fingerprints among N randomly selected individuals. Their equation (3.8) provides an excellent approximation based on $\Sigma\gamma_i^2$, $\Sigma\gamma_i^3$, and $\Sigma\gamma_i^4$, where all sums are over $i = 1,2, \ldots s$. Because all the γ_i are so small in the data we are considering, the simpler equation (3.10) in Gail et al. (1979) yields virtually identical results. It approximates the probability of two or more common fingerprints as

$$1 - \exp[-C(N,2)\ \Sigma\gamma_i^2] \,. \qquad \text{(A1)}$$

Equation (A1) is smallest, for fixed N, if all the γ_i are equal to s^{-1}. Thus, allocations of alleles to loci that result in a homogeneous set of probabilities $\{\gamma_i\}$ will tend to minimize the chance of one or more common fingerprints among N individuals. Likewise, allocation of alleles to loci in such a way as to cause some γ_i to be much larger than others will produce a greater chance of one or more common fingerprints. Therefore we chose two extreme hypothetical allocations of alleles to loci designed to produce relatively homogeneous $\{\gamma_i\}$ or very heterogeneous $\{\gamma_i\}$, in an effort to bound the true probability of one or more common signatures.

To achieve heterogeneous $\{\gamma_i\}$, we ranked the p_i with $p(1) > p(2) > p(3), \ldots p(121)$. Then $p(1)$ was assigned to locus 1, $p(2)$ to locus 2, $\ldots$ $p(10)$ to locus 10, $p(11)$ to locus 10, $p(12)$ to locus 9, and so forth until all $p(i)$ had been assigned, subject to the constraints that the sum of allele probabilities must not exceed 1.0 at any locus. Minor rounding errors were handled by renormalization of allele probabilities at each locus after allocation of alleles. This allocation is called model II in the text; the abundant alleles are evenly distributed over loci. This allele allocation produces great heterogeneity in γ_i. To compute the smallest γ_i under model II, we examined each locus in turn. For locus l with A_l alleles, we consider all $A_l(A_l+1)/2$ possible homozygous and heterozygous allele patterns, with probabilities P_{l1}^2, $P_{l2}^2 \ldots$, $2P_{l1}P_{l2}$, $2P_{l2}^2P_{l3}$, $\ldots$, $2P_{l2}P_{l3}$, and so forth. We found the smallest of these probabilities for each locus, l. Multiplying L such minima together, we

obtained the smallest possible value of γ_i under model II, namely, 1.8×10^{-35}. Likewise, the most probable composite band pattern was found by selecting the largest of the previous probabilities at each locus and multiplying these together to obtain 2.1×10^{-11}, which is more than 24 orders of magnitude greater than the minimum value above. This justifies our description of model II as leading to "heterogeneous" values of γ_i.

Relatively homogenous $\{\gamma_i\}$ were produced by filling up locus 1 with the largest $p(i)$ possible, then filling up locus 2 with the remaining largest $p(i)$ and so forth. Loci 8–10 contain only small $p(i)$ values, whereas loci 1–3 contain mainly large $p(i)$ values. We call this allocation model I in the text; the abundant alleles are clustered in a few loci. Under model I, the smallest value of γ_i was found to be 1.5×10^{-30}, and the largest value was 6.4×10^{-16}. Although these probabilities are vastly different, they differ by only 14 orders of magnitude, compared with 24 orders of magnitude in the "heterogeneous case" above. We therefore describe model I as leading to relatively "homogeneous" fingerprint probabilities.

In order to use equation (A1), we need to calculate $\Sigma\gamma_i^2$. For the homogeneous case (model I) there are $s = 0.158 \times 10^{18}$ different fingerprints, whereas for the heterogeneous case (model II), $s = 0.798 \times 10^{19}$. Even with modern computers it is not feasible to calculate this sum directly. Therefore, $\Sigma\gamma_i^2$ was estimated by the following Monte Carlo method. The $C(A_l + 1,2)$ genotypes were listed at locus l, $l = 1,2, \ldots 10$. One of these genotypes was selected at random with equal probability and independently at each locus. The corresponding γ_i were computed as the product of terms such as P_{lj}^2 or such as $2P_{lj}P_{lk}$, depending on whether the selected genotype at locus l was homozygous or heterozygous. Once γ_i was computed, so was γ_i^2, γ_i^3, and γ_i^4. This process was repeated 1 million times, and the average values $\overline{\gamma_i^2}$, $\overline{\gamma_i^3}$, and $\overline{\gamma_i^4}$ were computed. Then $\Sigma\gamma_i^2$, $\Sigma\gamma_i^3$, *and* $\Sigma\gamma_i^4$ were estimated as $s\overline{\gamma_i^2}$, $s\overline{\gamma_i^3}$, and $s\overline{\gamma_i^4}$, respectively. These quantities were substituted in Gail et al.'s equation (3.8) to produce the results in table 2. The simpler formula (A1) yields identical results in table 2.

Models with 11 Loci and Invisible Fragments for HaeIII

On the basis of "high resolution" experiments, we consider a model with 11 loci, of which 1.7 correspond to 11 alleles whose fragments are not routinely detectable. We were not able to calculate probabilities of DNA fingerprints without assuming which "visible" and "invisible" alleles were on specific loci. We therefore consider the following two specific models in which the 11 "invisible" alleles each have probability $r = (11-9.406)/11 = 0.145$ and in which one such allele resides on each locus. Models I and II are created by allocating the "visible" alleles to the 11 loci as described above. To compute the probability of a given "visible" DNA fingerprint, we assume independent assortment at each locus, as before, and for locus l we use terms such as r^2, $P_{lj} + 2P_{lj}r$, or $2P_{lj}P_{lk}$, respectively, according to whether zero, one, or two visible fragments are associated with that locus. Note that a single visible fragment may represent either a homozygous allele or a heterozygous visible-invisible combination. Calculations of the probabilities of two or more common DNA fingerprints in a random sample of N individuals are similarly modified to account for invisible fragments and 11 loci. In particular, there are now $s = \Pi_{l=1}^{11}\{C(A_l+1,2)+1\}$ possible visible DNA fingerprints, because some loci may yield no visible fragments, and the Monte Carlo estimation procedures must be modified to select from the $C(A_l+1,2) + 1$ distinguishable visible patterns at lo-

Table A1

Estimated Probability of Chance Identity of Two DNA Fingerprints in a Population of *N* Individuals

	Model I (homogeneous)	Model II (heterogeneous)
DNA fingerprint probability:		
Median	6.4×10^{-18}	5.7×10^{-18}
Range	1.1×10^{-20} to 2.3×10^{-16}	4.1×10^{-20} to 2.7×10^{-15}
Probability of two or more common DNA fingerprints:		
$N = 10^4$	$<10^{-7}$	$<10^{-7}$
$N = 10^5$	5×10^{-7}	6×10^{-7}
$N = 10^6$	5×10^{-5}	6×10^{-5}
$N = 10^7$	5×10^{-3}	6×10^{-3}

cus l by generating probabilities r^2, $2P_{lj}P_{lk}$ for all $j \neq k$ and $P_{lj}^2 = 2P_{lj}r$ for all j + 1,2, . . . A_l "visible" alleles.

The results of these calculations are summarized in table A1. It is seen that median DNA fingerprint probabilities are smaller by a factor of about eight, compared with estimates from the simple model with 10 loci (table 2). For model I the estimated probability of two or more common DNA fingerprints is nearly the same as for the simple model with 10 loci (table 5), whereas for model II the estimates of common DNA fingerprints are smaller by a factor of 10 than are estimates based on only 10 loci. We conclude that the simple model with 10 loci tends to yield conservative results.

References

Bell GI, Selby MJ, Rutter WJ (1982) The highly polymorphic region near the human insulin gene is composed of simple tandemly repeating sequences. Nature 295:31–35

Burke T, Bruford MW (1987) DNA fingerprinting in birds. Nature 327:149–152

Church GM, Gilbert E (1984) Genomic sequencing. Proc Natl Acad Sci USA 81:1991–1995

Collins T, Lapierre LA, Fiers W, Strominger JL, Packer JL (1986) Loss of HLA in culture. Proc Natl Acad Sci USA 83:446–450

Ferrone S, Pellegrino MA, Reisfeld RA (1971) A rapid method for direct HL-A typing of cultured lymphoid cells. J Immunol 107:613–615

Gail MH, Weiss GH, Mantel N, O'Brien SJ (1979) A solution to the generalized birthday problem with application to allozyme screenings for cell culture contamination. J Appl Prob 16:242–251

Gartler SM (1967) Genetic markers as tracers in cell culture. Second Decennial Review Conference on Cell Tissue and Organ Culture. Natl Cancer Inst Monogr 26:167–195

——— (1968) Apparent HeLa cell contamination of human heteroploid cell lines. Nature 217:750–751

Gey GO (1954–55) Some aspects of the constitution and behavior of normal and malignant cells maintained in continuous culture. In: The Harvey Lectures: Series L. Academic Press, New York, pp 154–229

Gill P, Jeffreys AJ, Werrett DJ (1985) Forensic application of DNA "fingerprints." Nature 318:577–579

Gold M (ed) (1986) A conspiracy of cells. State University of New York Press, Albany

Harris NL, Gang DL, Quay SC, Poppema SW, Nelson-Rees A, O'Brien SJ (1981) Contamination of Hodgkin's disease cell cultures. Nature 289:354–356

Hay R, Macy M, Chen TR, McClintock P, Reid Y (1988) American Type Culture Collection catalogue of cell lines and hybridomas, 6th ed. American Type Culture Collection, Rockville, MD

Hsu SH, Schachter BZ, Delanez NL, Miler TB, McKusick VA, Kennett RH, Bodmer JG, et al (1976) Genetic characteristics of the HeLa cell. Science 191:392–394

Jeffreys AJ, Brookfield JFY, Semeonoff R (1985*a*) Positive identification of an immigration test-case using human DNA fingerprints. Nature 317:818–819

Jeffreys AJ, Morton DB (1987) DNA fingerprints of dogs and cats. Anim Genet 18:1–15

Jeffreys AJ, Wilson V, Thein SL (1985*b*) Hypervariable minisatellite regions in human DNA. Nature 314:67–73

——— (1985*c*) Individual-specific fingerprints of human DNA. Nature 316:76–79

Jones HW Jr, McKusick VA, Harper PS, Wuu KD (1971) George Otto Gey: the HeLa cell and a reappraisal of its origin. Obstet Gynecol 38:945–949

Lander ES (1989) DNA fingerprinting on trial. Nature 339:501–505

Lavappa KS, Macy ML, Shannon JE (1976) Examination of ATCC stocks for HeLa marker chromosomes in human cell lines. Nature 259:211–213

Lynch M (1988) Estimation of relatedness by DNA fingerprinting. Mol Biol Evol 5:584–599

Maniatis T, Fritsch EE, Sambrook J (1982) Molecular Cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

Mann DL, Gilbert DA, Reid YA, Popovic M, Read-Connole E, Gallo RC, Gazdar AF, et al (1989) Origin of the HIV-susceptible human CD4+ cell line H9. AIDS Res 5: 253–255

Mann DL, Popovic M, Sarin P, Murray C, Reitz MS, Strong DM, Haynes BF, et al (1983) Cell lines producing human T-cell lymphoma virus showed altered HLA expression. Nature 305:58–60

Nakamura Y, Leppert M, O'Connell P, Wolff R, Holm T, Culver M, Martin C, et al (1987) Variable number of tandem repeat (VNTR) markers for human gene mapping. Science 235:1616–1622

Nelson-Rees WA, Daniels DW, Flandermeyer RR (1981) Cross-contamination of cells in culture. Science 212:446–452

Nelson-Rees WA, Flandermeyer RR (1976) HeLa cultures defined. Science 191:96–98

Nelson-Rees WA, Flandermeyer RR, Hawthorne PK (1974) Banded marker chromosomes as indicators of intraspecies cellular contamination. Science 184:1093–1096

Nelson-Rees WA, Hunter L, Darlington GJ, O'Brien SJ (1980) Characteristics of HeLa strains: permanent vs. variable features. Cytogenet Cell Genet 27:216–231

O'Brien SJ, Kleiner G, Olson R, Shannon J (1977) Enzyme polymorphisms as genetic signatures in human cell cultures. Science 195:1345–1348

O'Brien SJ, Shannon JE, Gail MH (1980) A molecular approach to the identification and individualization of human and animal cells in culture: isozyme and allozyme genetic signatures. In Vitro 16:119–135

Proudfoot NJ, Gil A, Maniatis T (1982) The structure of the human zeta-globin gene and a closely linked, nearly identical pseudogene. Cell 31:553–563

Reed KC, Mann DA (1985) Rapid transfer of DNA from agarose gels to nylon membranes. Nucleic Acids Res 13: 7207–7221

Reeders ST, Breuning MH, Davies KE, Nicholls RD, Jarman AP, Higgs DR, Pearson PL, et al (1985) A highly polymorphic DNA marker linked to adult polycystic kidney disease on chromosome 16. Nature 317:542–544

Schweder T, Spjotvoll E (1982) Plots of P-values to evaluate many tests simultaneously. Biometrika 69:493–502

Stoker NG, Cheah KSE, Griffin JR, Solomon E (1985) A highly polymorphic region 3′ to the human type II collagen gene. Nucleic Acids Res 13:4613–4622

Thacker JM, Webb B, Debenham PG (1988) Fingerprinting cell lines: use of human hypervariable DNA probes to characterize mammalian cell cultures. Somatic Cell Mol Genet 14:519–525

van Helden PD, Wiid IJ, Albrecht CF, Theron E, Thornley AL, Hoal–van Helden EG (1988) Cross-contamination of human esophageal squamous carcinoma cell lines detected by DNA fingerprinting analysis. Cancer Res 48:5660–5662

Wetton JH, Royston EC, Parkin DT, Walters D (1987) Demographic study of a wild house sparrow population by DNA fingerprinting. Nature 327:147–149

Wong Z, Wilson V, Jeffreys A, Thein S (1986) Cloning a selected fragment from a human DNA "fingerprint": isolation of an extremely polymorphic minisatellite. Nucleic Acids Res 14:4605–4616

Wyman A, White R (1980) A highly polymorphic locus in human DNA. Proc Natl Acad Sci USA 77:6754–6758